I've spent the month of April blogging my way through the tidyverse, while using my reading dataset from 2019 as the example. Today, I thought I'd bring many of those analyses and data manipulation techniques together to do a post about my reading habits for the year.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse
1.3.0 --
```

```
##  ggplot2 3.2.1      purrr   0.3.3
##  tibble  2.1.3      dplyr   0.8.3
##  tidyr   1.0.0      stringr 1.4.0
##  readr   1.3.1      forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
reads2019 <- read_csv("~/Downloads/Blogging A to Z/SaraReads2019_allchanges.
csv",
                      col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   Title = col_character(),
##   Pages = col_double(),
##   date_started = col_character(),
##   date_read = col_character(),
##   Book.ID = col_double(),
##   Author = col_character(),
##   AdditionalAuthors = col_character(),
##   AverageRating = col_double(),
##   OriginalPublicationYear = col_double(),
##   read_time = col_double(),
##   MyRating = col_double(),
##   Gender = col_double(),
##   Fiction = col_double(),
##   Childrens = col_double(),
##   Fantasy = col_double(),
##   SciFi = col_double(),
##   Mystery = col_double(),
##   SelfHelp = col_double()
## )
```

As you recall, I read 87 books last year, by 42 different authors.

```
reads2019 %>%
  summarise(Books = n(),
            Authors = n_distinct(Author))
```

```
## # A tibble: 1 x 2
##   Books Authors
##
## 1    87      42
```

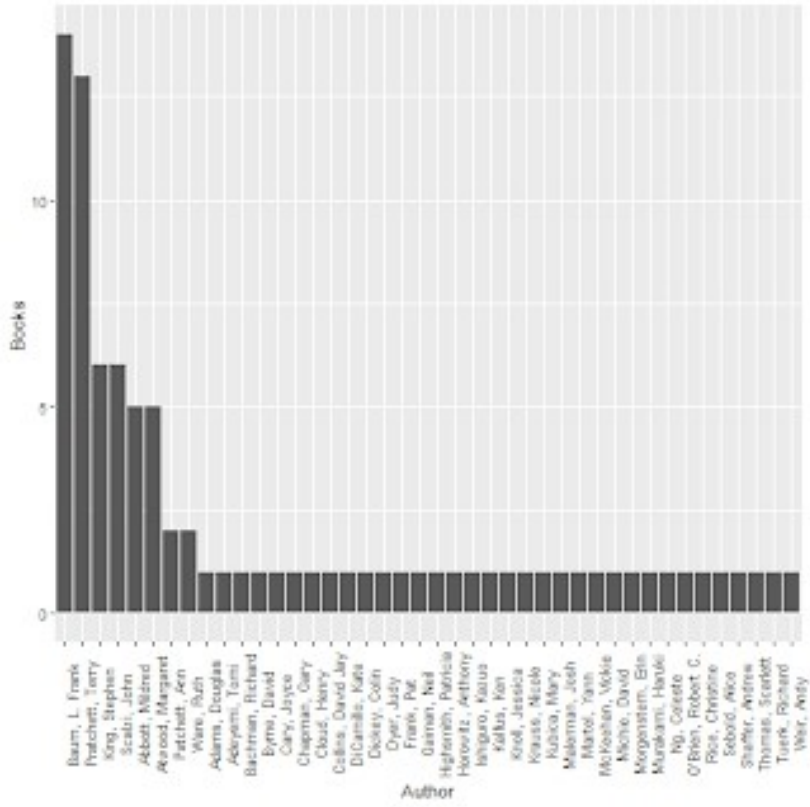
Using summarise, we can get some basic information about each author.

```
authors <- reads2019 %>%
```

```
  group_by(Author) %>%
  summarise(Books = n(),
            Pages = sum(Pages),
            AvgRating = mean(MyRating),
            Oldest = min(OriginalPublicationYear),
            Newest = max(OriginalPublicationYear),
            AvgRT = mean(read_time),
            Gender = first(Gender),
            Fiction = sum(Fiction),
            Childrens = sum(Childrens),
            Fantasy = sum(Fantasy),
            Sci = sum(SciFi),
            Mystery = sum(Mystery))
```

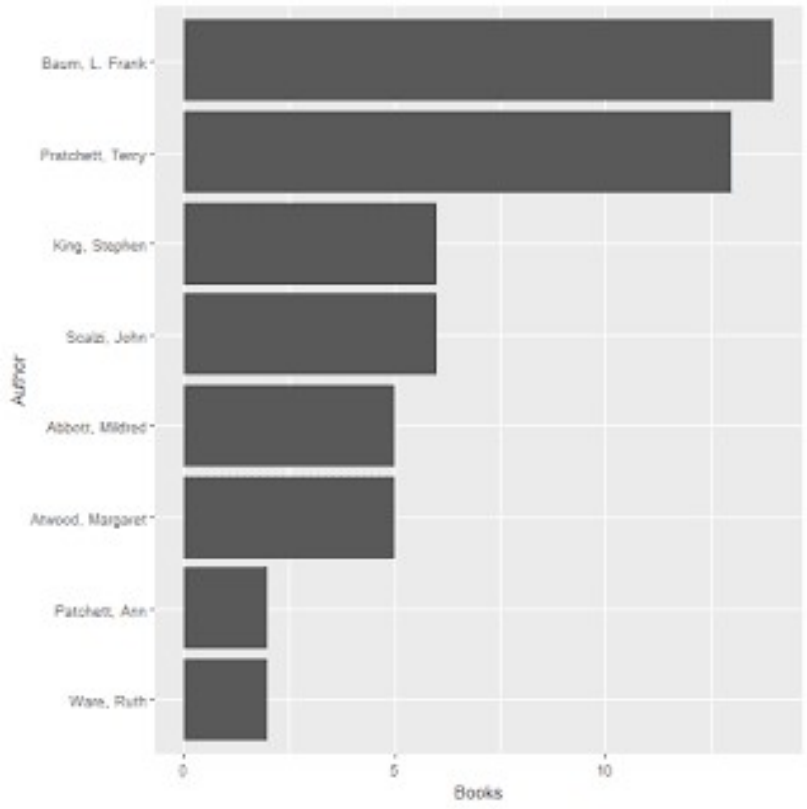
Let's plot number of books by each author, with the bars arranged by number of books.

```
authors %>%
  ggplot(aes(reorder(Author, desc(Books)), Books)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlabel("Author")
```



I could simplify this chart quite a bit by only showing authors with 2 or more books in the set, and also by flipping the axes so author can be read along the side.

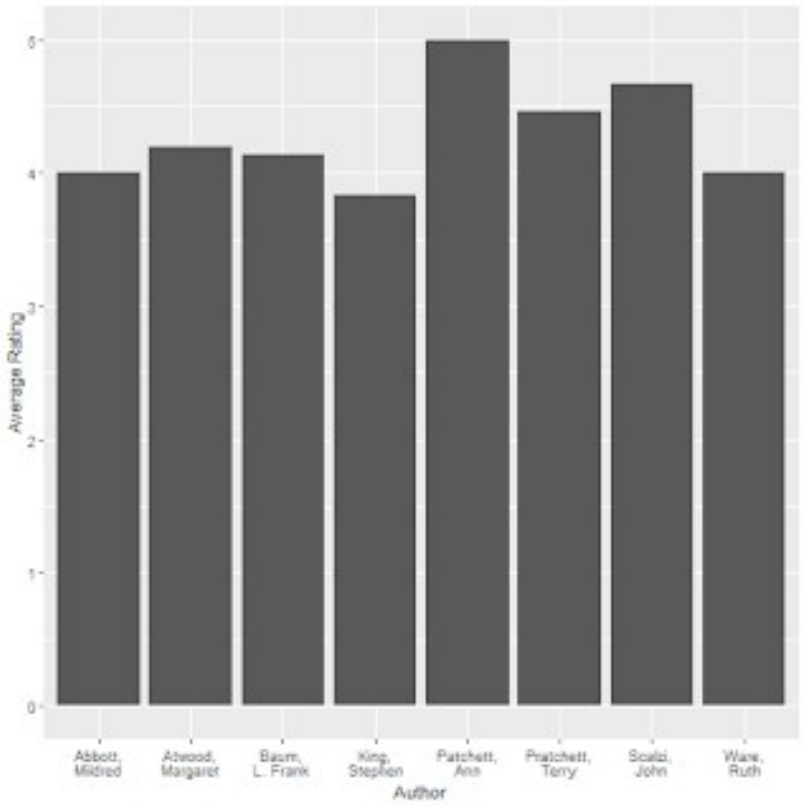
```
authors %>%
  mutate(Author = fct_reorder(Author, desc(Author))) %>%
  filter(Books > 1) %>%
  ggplot(aes(reorder(Author, Books), Books)) +
  geom_col() +
  coord_flip() +
  xlab("Author")
```



Based on this data, I read the most books by L. Frank Baum (which makes sense, because I made a goal to reread all 14 Oz series books), followed by Terry Pratchett (which makes sense, because I love him). The code above is slightly more complex, because when I use coord_flip(), the author names were displayed in reverse alphabetical order. Using the factor reorder code plus the reorder in ggplot allowed me to display the chart in order by number of books then by author alphabetical order.

We can also plot average rating by author, which can tell me a little more about how much I like particular authors. Let's plot those for any author who contributed at least 2 books to my dataset.

```
authors %>%
  filter(Books > 1) %>%
  ggplot(aes(Author, AvgRating)) +
  geom_col() +
  scale_x_discrete(labels=function(x){sub("\\s", "\n", x)}) +
  ylab("Average Rating")
```

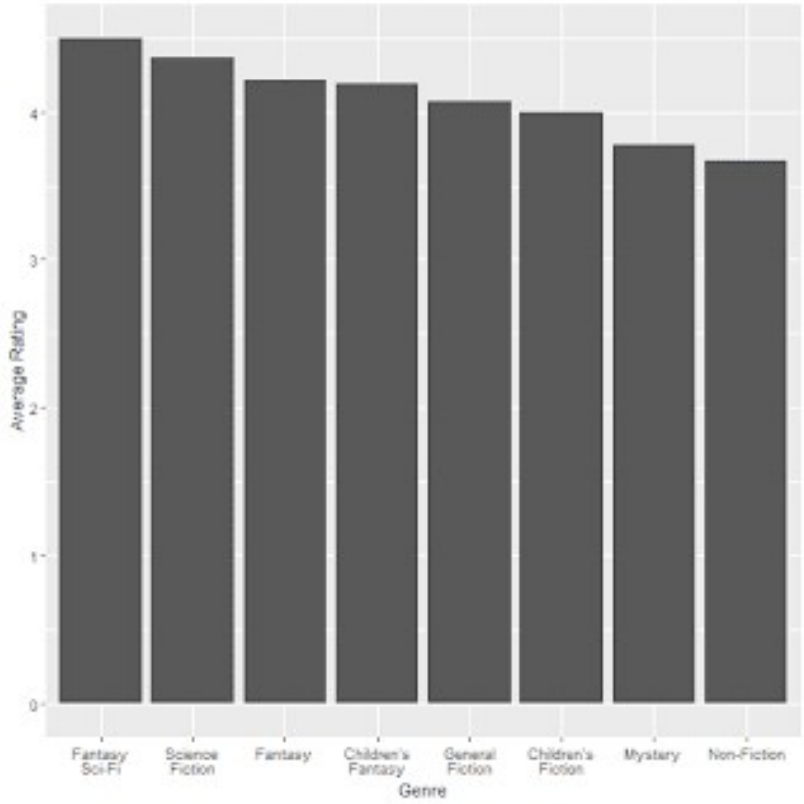


I only read 2 books by Ann Patchett, but I rated both of her books as 5, giving her the highest average rating. If I look at one of the authors who contributed more than 2 books, John Scalzi (tied for 3rd most read in 2019) has the highest rating, followed by Terry Pratchett (2nd most read). Obviously, though, I really like any of the authors I read at least 2 books from, because they all have fairly high average ratings. Stephen King is the only one with an average below 4, and that's only because I read *Cujo*, which I hated (more on that later on in this post).

We can also look at how genre affected ratings. Using the genre labels I generated before, let's plot average rating.

```
genre <- reads2019 %>%
  group_by(Fiction, Childrens, Fantasy, SciFi, Mystery) %>%
  summarise(Books = n(),
            AvgRating = mean(MyRating)) %>%
  bind_cols(Genre = c("Non-Fiction",
           "General Fiction",
           "Mystery",
           "Science Fiction",
           "Fantasy",
           "Fantasy Sci-Fi",
           "Children's Fiction",
           "Children's Fantasy"))

genre %>%
  ggplot(aes(reorder(Genre, desc(AvgRating)), AvgRating)) +
  geom_col() +
  scale_x_discrete(labels=function(x){sub("\\s", "\n", x)}) +
  xlab("Genre") +
  ylab("Average Rating")
```



Based on this plot, my favorite genres appear to be fantasy, sci-fi, and especially books with elements of both. No surprises here.

Let's dig into ratings on individual books. In my filter post, I identified the 25 books I liked the most (i.e., gave them a 5-star rating). What about the books I disliked? The lowest rating I gave was a 2, but it's safe to say I hated those books. And I also probably didn't like the books I rated as 3.

```
lowratings <- reads2019 %>%
  filter(MyRating <= 3) %>%
  mutate(Rating = case_when(MyRating == 2 ~ "Hated",
                   MyRating == 3 ~ "Disliked")) %>%
  arrange(desc(MyRating), Author) %>%
  select(Title, Author, Rating)

library(expss)

##
## Attaching package: 'expss'

## The following objects are masked from 'package:stringr':
##
##     fixed, regex

## The following objects are masked from 'package:dplyr':
##
##     between, compute, contains, first, last, na_if, recode, vars

## The following objects are masked from 'package:purrr':
##
##     keep, modify, modify_if, transpose

## The following objects are masked from 'package:tidyr':
##
##     contains, nest

## The following object is masked from 'package:ggplot2':
```

```
##
##     vars

as.etable(lowratings, rownames_as_row_labels = FALSE)
```

Title	Author	Rating
The Scarecrow of Oz (Oz, #9)	Baum, L. Frank	Disliked
The Tin Woodman of Oz (Oz, #12)	Baum, L. Frank	Disliked
Herself Surprised	Cary, Joyce	Disliked
The 5 Love Languages: The Secret to Love That Lasts	Chapman, Gary	Disliked
Boundaries: When to Say Yes, How to Say No to Take Control of Your Life	Cloud, Henry	Disliked
Summerdale	Collins, David Jay	Disliked
When We Were Orphans	Ishiguro, Kazuo	Disliked
Bird Box (Bird Box, #1)	Malerman, Josh	Disliked
Oz in Perspective: Magic and Myth in the L. Frank Baum Books	Tuerk, Richard	Disliked
Cujo	King, Stephen	Hated
Just Evil (Evil Secrets Trilogy, #1)	McKeehan, Vickie	Hated

I'm a little surprised at some of this, because several books I rated as 3 I liked and only a few I legitimately didn't like. The 2 books I rated as 2 I really did hate, and probably should have rated as 1 instead. So based on my new understanding of how I've been using (misusing) those ratings, I'd probably update 3 ratings.

```
reads2019 <- reads2019 %>%
  mutate(MyRating = replace(MyRating,
                            MyRating == 2, 1),
         MyRating = replace(MyRating,
                            Title == "Herself Surprised", 2))

lowratings <- reads2019 %>%
  filter(MyRating <= 2) %>%
  mutate(Rating = case_when(MyRating == 1 ~ "Hated",
                   MyRating == 2 ~ "Disliked")) %>%
  arrange(desc(MyRating), Author) %>%
  select(Title, Author, Rating)

library(expss)

as.etable(lowratings, rownames_as_row_labels = FALSE)
```

Title	Author	Rating
Herself Surprised	Cary, Joyce	Disliked
Cujo	King, Stephen	Hated
Just Evil (Evil Secrets Trilogy, #1)	McKeehan, Vickie	Hated

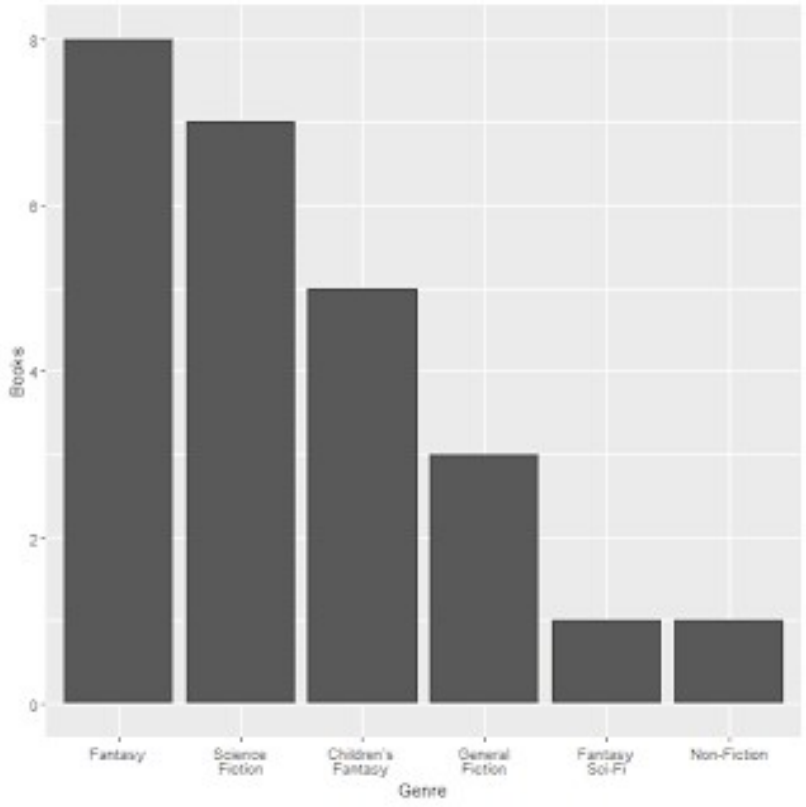
There! Now I have a much more accurate representation of the books I actually disliked/hated, and know how I should be rating books going forward to better reflect how I think of the categories. Of the two I hated, *Just Evil…* was an e-book I won in a Goodreads giveaway that I read on my phone when I didn't have a physical book with me: convoluted storyline, problematic romantic relationships, and a main character who talked about how much her dog was her baby, and yet the dog was forgotten half the time (even left alone for long periods of time while she was off having her problematic relationship) except when the dog's reaction or protection became important to the storyline. The other, *Cujo*, I reviewed here; while I'm glad I read it, I have no desire to ever read it again.

Let's look again at my top books, but this time, classify them by long genre descriptions from above. I can get that information into my full reading dataset with a join, using the genre flags. Then I can plot the results from that dataset without having to summarize first.

```
topbygenre <- reads2019 %>%
```

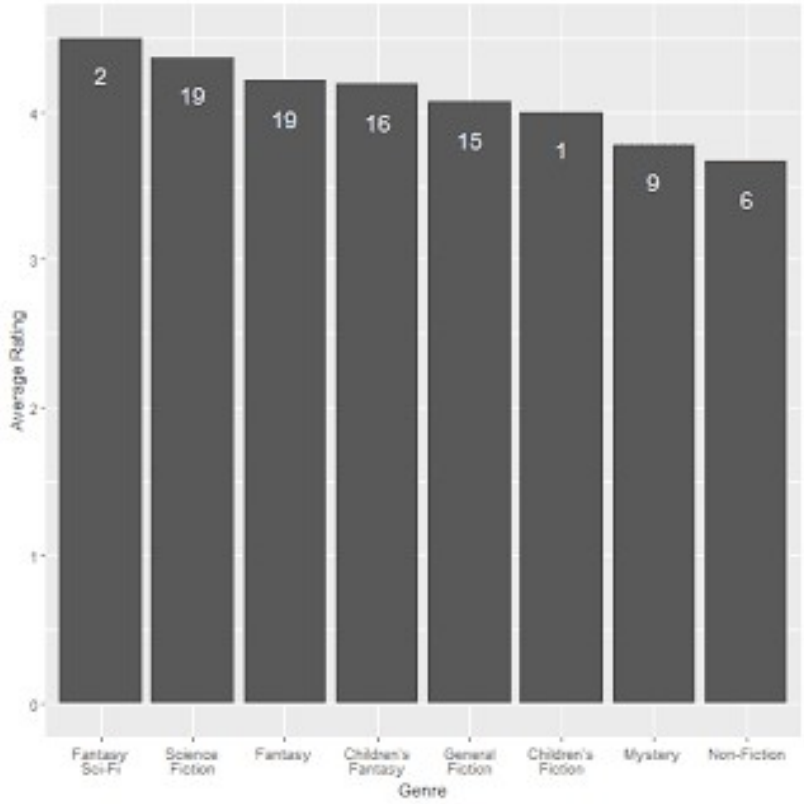
```
  left_join(genre, by = c("Fiction","Childrens","Fantasy","SciFi","Mystery"))
%>%
  select(-Books, -AvgRating) %>%
  filter(MyRating == 5)

topbygenre %>%
  ggplot(aes(fct_infreq(Genre))) +
  geom_bar() +
  scale_x_discrete(labels=function(x){sub("\\s", "\n", x)}) +
  xlab("Genre") +
  ylab("Books")
```



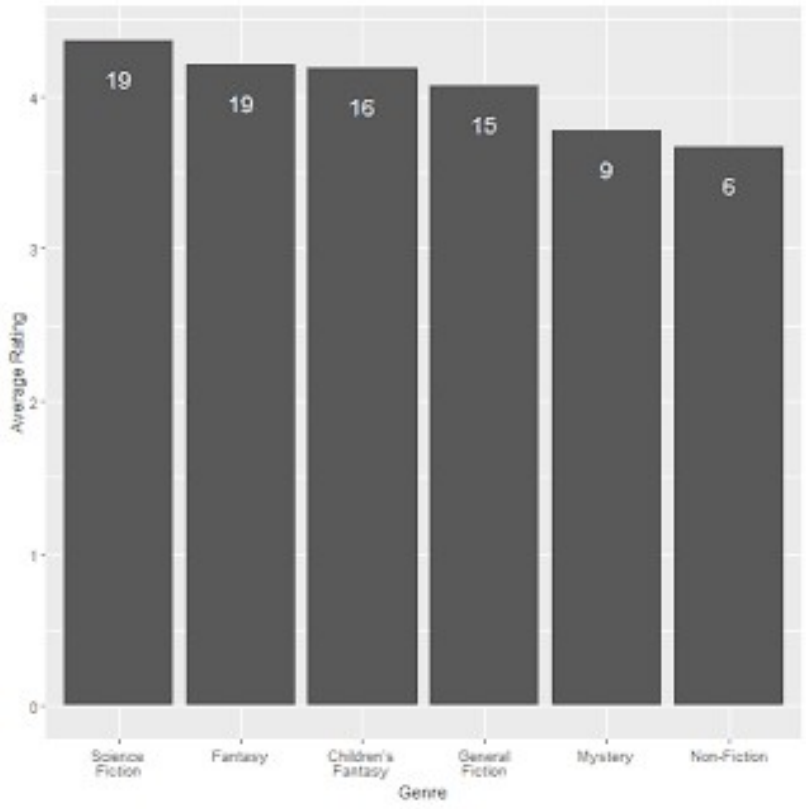
This chart helps me to better understand my average rating by genre chart above. Only 1 book with elements of both fantasy and sci-fi was rated as a 5, and the average rating is 4.5, meaning there's only 1 other book in that category that had to be rated as a 4. It might be a good idea to either filter my genre rating table to categories with more than 1 book, *or* add the counts as labels to that plot. Let's try the latter.

```
genre %>%
  ggplot(aes(reorder(Genre, desc(AvgRating)), AvgRating, label = Books)) +
  geom_col() +
  scale_x_discrete(labels=function(x){sub("\\s", "\n", x)}) +
  xlab("Genre") +
  ylab("Average Rating") +
  geom_text(aes(x = Genre, y = AvgRating-0.25), size = 5,
                color = "white")
```



Let's redo this chart, excluding those genres with only 1 or 2 books represented.

```
genre %>%
  filter(Books > 2) %>%
  ggplot(aes(reorder(Genre, desc(AvgRating)), AvgRating, label = Books)) +
  geom_col() +
  scale_x_discrete(labels=function(x){sub("\\s", "\n", x)}) +
  xlab("Genre") +
  ylab("Average Rating") +
  geom_text(aes(x = Genre, y = AvgRating-0.25), size = 5,
                color = "white")
```



While I love both science fiction and fantasy – reading equal numbers of books in those genres – I seem to like science fiction a bit more, based on the slightly higher average rating.